MANAGEMENT ANALYSIS & PLANNING, INC.

Review and Critique
Of

"An Evidence-Based Approach To School Finance Adequacy in Washington"
Draft dated June 28, 2006

James R. Smith

July 31, 2006

618 Del Oro Place, Davis, CA 95616 ♦ (530) 753-3130 ♦ map@edconsultants.com

Introduction

This paper addresses policy relevant issues associated with a report by Allan Odden, et al. dated June 28, 2006, and entitled "An Evidence-Based Approach to School Finance Adequacy in Washington." Odden and colleagues contend that Washington schools can dramatically improve student performance by adopting specified research-based interventions, and make recommendations for interventions and practices for which they claim there is "evidence" to support their implied promise of improved student performance.

Odden and his colleagues assert that it will be necessary for Washington to commit significant additional resources and reallocate existing resources to "adequately" fund its school finance system. Their report further implies that implementation of all of the consultants' recommendation will yield dramatic gains in student achievement. While Odden et al. immodestly entitle their methodology "evidence-based," they offer insufficient evidence that the implicitly promised outcomes will occur if their recommendations were implemented statewide. Virtually all of what they recommend is worthy of further investigation; moreover, probably little is likely to have any adverse affect on most students. But, it is important for policymakers to keep in mind that the consultants' recommendations are merely suggestive. The recommendations are not based on evidence that is sufficiently compelling to accept uncritically. Decisions on what to implement, when, and where are still appropriately made by the Legislature, Governor and other elected and appointed officials.

Their "evidence" is selective and not as compelling as a casual reader of their report may assume. The authors seem to accept uncritically studies that support their recommendations, and ignore studies that suggest different conclusions. They tend to accept studies that may not meet reasonable standards of scientific rigor, over-generalize from limited evidence, assume that a collection of interventions will be compatible in practice, they tend to not weigh the cost of interventions against likely outcomes, and assume that small-scale pilot projects and research studies can be generalized to statewide implementation. Each of these issues is addressed below. This paper is written for policymakers and as a consequence, footnotes and citations of related research have been minimized.

Studies cited may not meet standards of rigor.

Evidence offered in the report is from studies of interventions that may have been effective for some unspecified population of students in some unspecified context. Unless the research design of the studies upon which they rely met reasonable standards of scientific rigor, it is not possible to conclude with confidence that the interventions

actually produced the outcomes Odden and his colleagues assert. The report is largely silent about the nature and rigor of the various studies it cites.

Indeed the state-of-the-art of education research is such that this report, and similar reports that claim to measure the cost of an adequate education, are most appropriately viewed as advice, and only advice, to policymakers. The judgments and predilections of consultants should not trump the reasoned and nuanced judgments of legislatures and governors.

The US Department of Education What Works Clearinghouse[1] requires that any study rated as meeting evidence standards (with or without reservations) must employ one of the following research designs: a randomized controlled trial, a quasi-experiment with equating, a regression discontinuity design, or a single-case design. Wayne and Youngs[2] suggest that for studies to be considered compelling rather than merely suggestive they must account for students' prior achievement (gain scores) and control for student socioeconomic status. The consultants (Odden, et al.) should be required to report which of the studies they cite meet these standards.

The consultants over-generalize from limited evidence.

Recommendations based on "best practices" and the advice of interest groups are at best conventional wisdom. It is not evidence. Conventional wisdom in education is fragile and tends to shift with each new fad and wave of "reform." Open classrooms, team teaching, flexible scheduling, and New Math are but a few failed best practices that educators convinced policymakers to fund. Indeed corporal punishment was considered best practice until only a few decades ago.

Advice of professional associations is less than compelling as evidence. The counselors' association recommends a minimum caseload for counselors. The librarians' organization recommends librarians as essential to an adequately staffed school. But, where is the evidence that students learn more, or drop out less with the staffing patterns advocated by the interest groups? When one's only tool is a hammer, every problem strongly resembles a nail.

Even where research is cited, the consultants' report provides insufficient information to evaluate its utility. Before policymakers commit scarce resources to implement the consultants' recommendations, it would be responsible to require the consultants to conduct a thorough evaluation of each study or other source used to justify their recommendations. At the very least they should state their standards for selecting the studies upon which they rely. They also should report the design of each study, the number and demographic characteristics of subjects, how subjects were selected, the

[1] http://www.whatworks.ed.gov/

[2] Wayne, Andrew J. and Peter Youngs, "Teacher Characteristics and Student Achievement Gains: A Review," *Review of Educational Research, Spring 2003, Vol. 73, No. 1, pp 89-122*

qualifications of the educators producing the treatment, outcomes measured and how those outcomes were measured, and any sources of potential bias.

Policymakers often adopt a particular intervention or strategy even if the underlying evidence of effectiveness is ambiguous or lacking. Indeed uncertainty of outcome is common in social programs. But more and better information and evidence provides the opportunity for more enlightened decisions and potential for superior outcomes. More important, decision-making in an uncertain environment is more appropriately done by officials elected and accountable to voters than by consultants.

No evidence is offered that their model is coherent.

Even if the consultants had provided strong compelling evidence that every recommendation consistently produced improved student outcomes, they offer no evidence that each of the recommendations is essential or that all the recommendations are compatible. They imply that all are necessary. They offer no evidence of how one intervention affects another, but imply that the effects are additive. Aspirin reduces pain, Tylenol reduces pain, ibuprofen does too. But no responsible physician suggests taking the recommended dosage of all three will produce three times as much pain relief[3]. Nor would she suggest that all three are essential just because each one produces a positive outcome. It is merely conjecture that all of the consultants' recommendations operating together would produce outcomes superior to some subset of interventions or even a different set of interventions. It seems plausible that schools with classes as small as they recommend may not need or even productively employ specialists or tutors. It seems equally plausible the specialists could serve also as professional development coaches. Without investigating such questions, uncritically adopting the consultants' recommendations invites over-staffing and redundant functions. They have made nearly identical recommendations in several other states[4]. If there is evidence from those states, the consultants should cite it.

Cost effectiveness is not addressed.

The consultants set no priorities for which interventions may offer the highest payoff or the most "bang-for-the-buck." They appear to assume that any intervention associated with potentially improved outcomes should be implemented regardless of cost. This practice would make sense only in an environment of unlimited resources. For example, the consultants report an effect size of .25 for class size reduction, which is easily the costliest of their recommendations. Multi-age classes (which theoretically would cost nothing extra) are reported to produce an effect size double that of class size reduction. Imbedded technology promises an effect size of .30 to .38 for what must be a tiny fraction of the cost of class size reduction. But none of the recommendations come close

[3] This analogy was suggested by Professor Michael Podgursky, Department of Economics, University of Missouri

[4] Apparently without regard to differences in student demographics or state standards and policies.

to being as cost effective as professional development with an effect size of as much as 2.7. If policymakers believe this effect size, a prudent decision may be to adopt only this recommendation, as it promises that more than 95 percent of students will score proficient (assuming this was the outcome measured in the research).

The following hypothetical example suggests a more useful way for policymakers to evaluate the various recommendations. Suppose that the decision is focused on a single school. Further suppose that lowering class size by two students costs $50,000 and has an effect size of .4. Hiring teacher mentors costs $20,000 and has an effect size of .3. Clearly the achievement yield per thousand dollars spent on class size reduction (.4 / 50 = .008) is smaller than the yield per one thousand dollars spent on teacher mentors (.3 / 20 = .015).

Pilot projects may not generalize to statewide implementation.

Successful results from a single study do not necessarily mean that it can be replicated. Under the best of circumstances research based on small groups[5] of subjects can demonstrate that an intervention can produce successful outcomes, but it may not demonstrate that it will work under real world conditions. Frequently, if not usually, it is very difficult to translate pilot studies or research projects into large scale implementation, at least in part because incentives and other conditions for success may not obtain on a larger scale.

Incentives in public schools are diffuse. Producing academic achievement of students competes for resources and attention with employment, working conditions, sports, social development, etc. For example, Hoxby[6] examined naturally occurring reductions in class size, where there were no particular systematic changes in motivation or incentives perceived by participants, and found no effect of reduced class size on student achievement. Until the Tennessee STAR study, most evidence on class size reduction was mixed at best. It seems reasonable to speculate that the teachers participating in the STAR study were more focused on producing higher test scores if they assumed, for example, that future funding for smaller classes was conditioned on the success of the research project[7].

In 1998 the New Jersey Supreme Court mandated that that state implement in the Abbott[8] school districts preschool programs for three and four year olds, standards-based education, whole school reform (Success for All), new and rehabilitated facilities, and

---

[5] Commonly cited research projects of successful preschool programs were conducted with fewer than 200 subjects, and carefully chosen providers.

[6] Hoxby, Caroline M. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *The Quarterly Journal of Economics,* Nov. 2000, pp 1239-1285

[7] Arguably, the Hoxby study more closely replicates the conditions of statewide implementation than controlled studies.

[8] *Abbott v. Burke, 163 N.J. 95*

supplemental programs such as family support teams, social and health services, alternative education programs, summer school, after-school, improved parent participation, and other reforms. Further, the Court mandated that the preschool programs employ certified teachers and have class sizes no greater than 15. In spite of the Abbott Districts spending annually as much as $19,000 per pupil, student outcomes remain disappointing low[9].

In many cases it is difficult to replicate successful studies because the resources employed under study conditions may not be universally available. This may be particularly true when highly motivated and talented educators provide the studied intervention. The necessary levels of motivation and unique talent may not be available in all districts or it may be difficult for educators to sustain the requisite level of effort for extended periods of time.

Some interventions appear to affect students with specific demographic characteristics differentially. For example, evidence from the STAR study suggests that smaller classes were more effective for poor and minority students, the population that was disproportionately represented as research subjects. As a consequence, under any circumstances, lowering class size as suggested by the consultants is unlikely to yield the statewide effects they suggest (because most of the student population in WA is not poor and minority) and may produce little change in achievement in schools that are predominantly not poor and not minority[10]. Therefore, it is critical for the consultants to describe in detail the nature of the student subjects in each study underlying their recommendations so that policymakers can infer the likely outcome if the proposed intervention were to be implemented with various populations and communities statewide.

On pages 19-20 of their report, the consultants side with Hedges, Laine and Greenwald in the debate between those researchers and Hanushek over whether increases in funding are associated with improvements in student outcomes[11]. Hedges, et al. conclude that considering effect sizes, on average more funding improves student achievement. In other words some schools improved, others did not. Those that improved raised the overall average. Still, most studies did not demonstrate improvement with added resources. Policymakers need not take sides in this debate to view it as a possible caveat to be applied in the context of the Odden et al. recommendations. The consultants imply that adopting their recommendations will improve all schools. For all the reasons discussed

---

[9] Allan Odden served as a court appointed special master for the Abbott districts.

[10] The school size research also suggests that poor and minority students are the primary beneficiaries of smaller schools and that smaller schools are associated with little change in outcomes for students who are not poor and not minority.

[11] Subsequent research using gain score data has revealed the relationship between increased inputs and student outcomes to be even more tenuous. See Hanushek, Eric A., "The Failure of Input-Based Schooling Policies," *The Economic Journal, 113 (February)2003*

above, this is an unlikely scenario. It is more likely that only some schools will translate additional resources into improved student outcomes. Others may even implement all of the recommendations and still fail to improve student outcomes. The experience of reform in other jurisdictions suggests that the former are the schools already producing better student outcomes. The latter are more likely schools populated with poor and minority children.

School finance in Wyoming offers a cautionary tale. In that state per pupil expenditures increased from $5,971 in 1996-97 to an estimated $12,422 for 2006-07. In spite of dramatic increases in spending, NAEP scores have remained essentially flat[12]. Perhaps even more troubling is Wyoming's experience with special education. School districts are reimbursed 100 percent for special education expenditures. Even in light of virtually unrestricted resources to educate identified students, and the fact that more than 80 percent of special education students are mildly handicapped (speech and learning disabled), only 5 percent to 18 percent (depending on test and grade level) of all students with IEPs scored at least proficient on the state assessment in 2005[13].

Odden, et al. argue that whether more money makes a difference depends on how the money is spent, implying money spent on their recommendations will produce the promised outcomes; but possibly it is a question of how well money is spent, maybe even regardless of the interventions implemented.

Specific observations from the report.

The following are observations about specific evidence adduced by the consultants throughout their report.

Page 17, footnote 5. The proposal that all students be required to complete a college prep course of study is not based on research. The proposal is controversial and is more appropriately a policy decision. One-size-fits-all policies rarely work as intended.

Page 28-30, Class size recommendations. Class size reduction easily is the most expensive intervention recommended[14]. The consultants recommend class sizes of exactly 15 for grades K-3 based almost exclusively on the STAR study. However, the STAR study employed class sizes of 13-17. Even relying on STAR, without compelling evidence that classes of 15 are superior to those of 16 or 17, the fiscally prudent decision would be to implement class sizes of 17. At least two other points should be considered about the STAR study. First, most of the gain in student achievement was observed in students' first year in smaller classes. It may be more efficacious to lower class sizes for Kindergarten and perhaps first grade only. Second, the benefit of smaller classes was

[12] And generally lower than those of Montana which spends significantly less per pupil.
[13] https://wdesecure.k12.wy.us/stats/wde_public.esc.show
[14] In addition to the cost of adding staff, many districts would need to construct classrooms to accommodate the increase in the number of classes.

significantly greater for poor and minority students. Again, the more cost effective decision may be to reduce class size only for schools with large populations of poor and minority students.

The class size recommendations for grades 4-12 are not based on research. The professional judgment of educators or researchers should be taken into account and given the weight it deserves; but it is not research. It is not evidence. Therefore, there is no evidence to suggest that class sizes of 26, 27 or 28 would produce outcomes inferior to class sizes of 25. Statewide the cost difference would, however, be substantial.

Page 31. Specialist teachers. The consultants make an argument for specialist teachers, but do not cite evidence that specialists produce superior student outcomes. It would be useful to know if there is a class size at which specialists become redundant.

Page 32. Block scheduling. The consultants cite no evidence that block scheduling produces outcomes superior to more traditional schedules. It does cost more.

Page 33. Instructional facilitators. The consultants cite no evidence of changes in student outcomes for the specific staffing ratios they recommend. Instead, they cite themselves as having made similar recommendations in three other states. Citing one's published research may be acceptable evidence, but the fact that one consistently makes a recommendation is not. It is merely one's opinion repeated.

Page 34. Professional development. The reported effect sizes cited are implausible. If they can be substantiated, professional development would seem to be a cost effective intervention if qualified coaches could be hired or identified in the numbers necessary to staff schools statewide.

Page 39. Tutors. The recommendation of at least one tutor per school ignores differences in student populations and assumes that every school can productively employ a full time tutor. Moreover, tutors are layered on top of recommendations for small classes, full-day kindergarten and other resources for struggling students. This is an example of where policymakers have no evidence on whether the various interventions are compatible or if some may be redundant.

Page 41. Bilingual education. An extensive body of research, conducted earlier than the study cited, has found no particular advantage of bilingual education over other strategies for instructing ELL.

Page 42. Extended-day programs. The consultants admit that the evidence on extended-day programs is mixed. It would be useful to know if such programs are cost-effective in schools as richly resourced as those envisioned by the consultants.

Page 45. Summer school. Again the cited results are mixed and less than compelling.

Page 48. Alternative schools. The consultants provide no evidence of the effects of alternative schools.

Page 57. Vocational equipment. No justification is provided for the recommendation for $7000 per vocational teacher. More important is how do the consultants reconcile vocational programs with the requirement that all students complete a college prep program? In some schools where this is policy, the very students who are likely to benefit from vocational programs find little time for classes beyond the required academic classes and remedial classes necessary for them to master the college prep curriculum.

Pages 57-58. Student support/Family-community outreach. The consultants assert that schools need a student support and family outreach strategy, but offer no evidence of the effect of such a strategy or if it is necessary in all schools. Citing comprehensive school designs is not evidence.

Page 59. Guidance counselors. The American School Counselor Association is an advocacy organization for school counselors. It is unlikely that their recommendations for staffing ratios are based on research. If it is, the consultants should cite it.

Page 60. Librarians. While it is conventional wisdom that schools need a librarian, there is no evidence cited to support the consultants' recommendation.

Pages 61-62. Principals. Traditionally public schools employ a principal. The recommended staffing ratios are, however, arbitrary and not based on any cited evidence.

Page 64. Effect sizes. In education research it is common to report the impact of a particular factor on student achievement in the form of "effect sizes." For example, an effect size of .4 means that the difference in achievement between the "treated" and the "control" students is .4 standard deviations. Effect sizes allow us to compare the impact of different interventions in a common metric (standard deviations of student achievement). If the effect size of intervention A is .4 and that of intervention B is .8, we can conclude that the latter has twice the effect of the former. As discussed above, just comparing effect sizes does not lead to appropriate decisions about how resources should be allocated.

Odden, et al. report effect sizes, but they do not report the outcome measure of the cited study. One can not tell, from what they report, whether the score that improved was from a academic test, or was some other outcome such as attendance, graduation, attitudes, or self esteem. If it was a test, they do not report if it was an easy test, difficult test, or if it was a nationally published test or one designed by the researchers. It is important to know if the outcome was trivial or sufficiently important for policy makers to spend money to replicate it.

Conclusion and Recommendations

Odden et al. are asking Washington policymakers to make a large wager of state revenue that the interventions and expenditures they recommend will produce doubling and tripling of student performance. But, a cursory analysis of the Odden report reveals that the evidence upon which they base their recommendations is equivocal and in some instances nonexistent.

The consultants' report is perhaps most useful as a compendium of interventions that might work under certain circumstances. It may be prudent to selectively implement those interventions that show the most promise for Washington schools and systematically evaluate them in varied contexts. Keeping in mind the difficulties associated with scaling up from pilot studies, it is essential that the design of such studies meet accepted standards of scientific rigor. Under ideal conditions policymakers would choose to implement only those interventions that they are confident are likely to succeed statewide in Washington schools with Washington students. Unfortunately, conditions of perfect information are rare, and policymakers must make decisions based on less than perfect information. Legislatures, governors and other policy makers are the appropriate entities to make these decisions, using the best information that is available to them at any particular time.